Representing variation in a spoken corpus of an endangered dialect: the case of Torlak

Vuković, Teodora

Abstract: The paper presents a spoken corpus of the endangered Torlak dialect from the Timok area of Southeast Serbia. This dialect expresses a great deal of variation in the use of non-standard features under the influence of standard Serbian (SSr). Accounting for this variation, a specific methodology has been selected for collection, sampling, transcription and annotation. Between 2015 and 2017, semi-structured interviews were conducted in the field eliciting spontaneous speech in the form of long narratives about traditional culture and history. The corpus comprises 500,697 tokens of semi-orthographic transcripts representing 80 h of recording from locations evenly distributed across the Timok area of the Torlak dialect zone, thus enabling a spatial contrastive analysis. The majority of speakers in the corpus are older people whose language represents the highly non-standard variety. In order to allow for analysis of language change under the influence of SSr, the corpus includes a number of younger people whose speech is closer to SSr. Tools for automatic PoS annotation and lemmatization that were lacking were developed based on the existing resources for SSr. For tagger training, a dialect sample of 27,000 manually verified tokens was merged with an existing training set for SSr.

DOI: https://doi.org/10.1007/s10579-020-09522-4

Originally published at:
Vuković, Teodora (2021). Representing variation in a spoken corpus of an endangered dialect: the case of Torlak. Language Resources and Evaluation, 55(3):731-756.
DOI: https://doi.org/10.1007/s10579-020-09522-4

Lang Resources & Evaluation
https://doi.org/10.1007/s10579-020-09522-4

ORIGINAL PAPER

Representing variation in a spoken corpus of an endangered dialect: the case of Torlak

Teodora Vuković[1]

Accepted: 2 December 2020

Abstract The paper presents a spoken corpus of the endangered Torlak dialect from the Timok area of Southeast Serbia. This dialect expresses a great deal of variation in the use of non-standard features under the influence of standard Serbian (SSr). Accounting for this variation, a specific methodology has been selected for collection, sampling, transcription and annotation. Between 2015 and 2017, semi-structured interviews were conducted in the field eliciting spontaneous speech in the form of long narratives about traditional culture and history. The corpus comprises 500,697 tokens of semi-orthographic transcripts representing 80 h of recording from locations evenly distributed across the Timok area of the Torlak dialect zone, thus enabling a spatial contrastive analysis. The majority of speakers in the corpus are older people whose language represents the highly non-standard variety. In order to allow for analysis of language change under the influence of SSr, the corpus includes a number of younger people whose speech is closer to SSr. Tools for automatic PoS annotation and lemmatization that were lacking were developed based on the existing resources for SSr. For tagger training, a dialect sample of 27,000 manually verified tokens was merged with an existing training set for SSr.

Keywords Spoken corpora · Non-standard corpora · Part-of-speech annotation · Lemmatization · Manual annotation · Torlak · Serbian

✉ Teodora Vuković
teodora.vukovic2@uzh.ch

[1] Slavisches Seminar, University of Zurich, Plattenstrasse 43, 8032 Zurich, Switzerland

Published online: 09 January 2021

1 Introduction

The Torlak dialect is a group of South Slavic varieties spoken in Southeast Serbia, northern parts of Macedonia and western parts of Bulgaria.[1] The Torlak area encompasses the Prizren-Timok dialects in Serbia, for the largest part, as well as the Kumanovo, Kratovo and Kriva Palanka dialects in Macedonia, and the Belogradchik, Godech, Tran, Breznik varieties in Bulgaria (Salminen 2010; Woolhiser 2011; Sobolev 1998), Fig. 1. Today, Torlak is officially listed as an endangered language by the UNESCO (Salminen 2010), partly due to the large decline in the population in Southeast Serbia, which covers the majority of the Torlak areas (Penev and Marinković 2012). Another reason is the dialect's low prestige in comparison with other forms of Serbian. Torlak is associated with a stereotypical image of low culture and education. As a consequence, speakers, especially young ones, often adopt the use of standard forms when confronted with speakers of standard Serbian (Petrović 2015). The non-standard features of Torlak are perishing under the influence of the standard as Vuković and Samardžić (2018), as well as others in Ćirković (2018) report.

Torlak is well described in the traditional dialectological sense—its non-standard features have been listed and defined, starting from Belić (1905) and Stanojević (1911), and repeated in a similar fashion onwards (Alexander 1975; Ivić 1985; Dinić 2008; Bogdanović 2015). In the present situation these descriptions must be considered incomplete as they do not account for the diffusion of the dialect's features, the variation in the context of the South Slavic dialect continuum, and the effects of the circumstances that have brought Torlak to its present endangered status.

One reason is that the linguistic data on Torlak is lacking. Texts and transcripts in the dialect have been published several times throughout the twentieth century in printed form, but never digitally. A publicly available electronic resource would provide for the study of the dialect itself and comparison with neighbouring languages and dialects. The necessity for such corpora has been recognized and addressed in surrounding countries, for Bulgarian, Croatian and Slovene, a fact which highlights the importance of contributing data for spoken as well as dialectal Serbian. In order to fill this gap, we are building a corpus of the Torlak dialect as spoken in Southeast Serbia based on recordings of fieldwork interviews collected in the Timok region between 2015 and 2018 (Fig. 1).[2]

Being the first resource of its kind, there are no automatic processing tools that could be used for Torlak. In developing the tools for part-of-speech (PoS) annotation and lemmatization, we are building upon existing tools for standard Serbian (SSr) as the closest resourced language. The tools are trained based on a manually verified training dialect set of 27,000 tokens.

[1] By using the term Torlak alone or together with terms such as dialect, language or other synonyms, we are not making any implications to its political status.

[2] Maps were created using Leaflet package in R (Graul 2016). The base map is Stamen TonerLite (https://stamen.com).

The paper covers collecting, sampling and transcription of the data, as well as the development of tools for PoS annotation and lemmatization. The resulting corpus documents an endangered under-resourced variety and provides valuable research material.

2 Torlak dialect

Although classified as a dialect of Serbian, Torlak shows considerable differences from standard Serbian (SSr) at phonological, lexical, morphological and syntactic levels (Belić 1905; Stanojević 1911; Ivić 1985; Alexander 1975), sharing many features with Bulgarian and Macedonian (Salminen 2010; Ivić 1985; Lindstedt 2000). The most non-standard varieties of Torlak can sometimes be unintelligible to speakers of SSr. Like any regional variety, it has numerous lexemes that are either archaic, such as *oganj*, or that are not a part of the SSr lexicon, such as *preripi* (Example 1). Phonologically, the most noticeable difference is in accent position and quality (Examples 1 and 2). Furthermore, there are genealogical differences regarding the reflection of Old Church Slavonic *yat* as in the *nee – nije* distinction (Example 2) or the retention of the syllable-final *l* /l/, which has changed into *o* /o/ in SSr, e.g., *oćal* compared to *hteo* in SSr (Example 2); the use of the central vowel *ə* / *ə*/ in some contexts where *a* would be used in SSr, as in *səg* (Example 2) or the palatal *ќ* /c/ and *cveќe* (Example 2), instead of the velar *k* /k/, in SSr. Another phonological difference is the complete or contextually dependent absence of some phonemes, like the fricative *h* /x/ in the distinction *reko – rekoh* or the approximant *j* in *edna* (Example 1). Some morphological differences regarding verbal inflection are manifested in the different affixes in *povalj-u* and *popad-a-ju* (Example 1). Morphosyntactic features include the use of the post-posed demonstrative clitic as in *oganj-at* (Example 1). Syntactic characteristics involve omission of the auxiliary or complementizer (Example 2).

1) Tor *Oná* *[ide]* *od* *ovúdaka* *i* *prerípi*
she.SG.NOM go.3SG.PRES from here and jump.3SG.PRES
ogánj-at, *oné* *žené* *se* poválju
fire.M.SG.ACC-DEM those.F.NOM woman.F.PL.NOM REFL roll.3PL.PRES
jedná *tám* *edná* *vám.*
one.F.SG.NOM there one.F.SG.NOM here

SSr *Òna* *[ide]* *odàvde* *i* *prèskoči*
she.SG.NOM go.3SG.PRES from here and jump.3SG.PRES
vàtru, *òne* *žène* *pòpadaju*
fire.M.SG.ACC-DEM those.F.NOM woman.F.PL.NOM fall.3PL.PRES
jèdna *tàmo* *jèdna* *ovámo.*
one.F.SG.NOM there one.F.SG.NOM here

'She goes from here and jumps over the fire, those women fall over, each to a different side.'

2) Tor *Oćál* *ón* *ponesé* *cveće* […]
want.3SG.PPART he.SG.NOM bring.3SG.PRES flower.M.PL.ACC
née *kako* *sȁg.*
to be.3SG.PRES.NEG as now

SSr Htêo je ón da ponèse cvêće
want.3SG.PPART AUX.3SG he.SG.NOM COMP bring.3SG.PRES flower.M.PL.ACC
[…] nìje kao säda.
to be.3SG.PRES.NEG as now
'He would bring flowers, not like today.'

2.1 Variation in Torlak

Like any linguistic area, the Torlak dialect zone is not uniform, but rather shows diatopic and diastratic variation. This information is not visible in the current descriptions of the dialect. Sobolev (1998), for instance, presented the horizontal distribution and isoglosses of many relevant features in his Atlas. His collection and the proceeding descriptions are certainly an admirable feat, but they do not show the degree of variation in each location. The current descriptions show a binary distinction—whether a feature is used or not. This is a useful distinction to some extent: not all features are used in the entire area. But a closer look at the speech samples reveals that intra-speaker variation exists in almost every individual and that most features are used in ways that are inconsistent and nuanced.

This variation is further influenced by the prestige of the standard languages of the area. What was a unified area until the nineteenth century, has been divided by the national boundaries of Serbia, Bulgaria and Macedonia, and the dialect has been converging towards the standard of each of these countries. At present, it is often restricted to use within the home or in a familiar environment (Salminen 2010). Over the course of fieldwork, it was easy to notice that older people use more dialectal features in comparison with the younger population (Vuković and Samardžić 2018; Ćirković 2018).

This variation has not been examined mainly because no data have been available. Looking into the variation patterns can give insight into the interplay of the languages in the area, but also about linguistic variation in general. Parallel patterns exist everywhere in the corpus as speakers can often switch between standard and non-standard forms even within a single sentence, as in the case of *edan* and *jedan* (1a), for instance. Correlating grammatical observations with extralinguistic factors, demographic or geographic information can bring to light the drivers of change beyond the language itself. For instance, a recent study of the frequency of post-positive demonstratives shows that they are used more frequently in isolated locations than in places closer to urban centres (Vuković and Samardžić 2018).

By creating a corpus, we aim to capture this manifold variation that can be observed in spoken Torlak. This does not come without challenges, of course. One challenge is to provide a representative sample of the dialect, including its non-standard features and especially their variation. This involves a methodology of collecting naturally occurring language, selecting a sample that represents how the

dialect is used in the whole Timok area, as well as a transcription methodology that gives an accurate illustration of dialect features. Another challenge is to create a PoS tagger and the other processing tools required to be able to identify different variants of a word and assign them a corresponding tag with as much accuracy as possible. In the following chapters, we describe our approach and the resulting resources.

3 Related work

As opposed to written language, spoken language exhibits arguably more spontaneous and natural language use, while non-standard varieties give an insight into changes that occur in language across space and time, as well as in contact with other languages or varieties. The corpora of spoken language and regional varieties are scarce compared to the resources for the written language, even in the case of major world languages, such as English or German (Anderwald and Wagner 2007; Anderson et al. 2007; Schmidt 2014; Samardžić et al. 2016), and even more so for languages such as Serbian. This is because creating such a resource is often a difficult and time-consuming process involving a lot of manual work to transcribe and process data before it takes on the form of a corpus, as many accounts in Beal et al. (2007) relate.

Outside of the South Slavic linguistics, there are many notable dialect corpora featuring spoken samples, which can be used as examples of a good practice. Here, we will mention some of them. The Nordic Dialect Corpus[3] consists of 2.75 million words of spontaneous speech and written samples from dialects of the North Germanic languages across all of the Nordic countries, including Norwegian, Swedish, Danish, Faroese, Icelandic and Övdalian languages. The transcripts are linked to audio and video, the corpus has a map function, and can be searched in a variety of ways, using syntactic, PoS annotations and lemmatization. The corpus is a general one, even though the original aim of the corpus was to focus on syntax (Johannessen et al. 2009, 2014, NDC Website). The ArchiMob[4] corpus of Swiss German dialects (Samardžić et al. 2016; Scherrer et al 2019) contains over 500,000 tokens of transcribed text intended for the analysis of the geographic distribution of morphosyntax and natural language processing. The corpus has been semi-automatically annotated with PoS tags. Due to the high variability and the lack of a written standard for Swiss German, the corpus has been semi-automatically normalized into Standard German. The Newcastle Electronic Corpus of Tyneside English (NECTE)[5] is a spoken dialect corpus from Tyneside in England. NECTE materials are stored in a Text Encoding Initiative (TEI)-conformant XML-encoded corpus and provide a variety of aligned formats: digitized audio, standard orthographic transcription, phonetic transcription, and part-of-speech annotation (Allen et al. 2007). Such rich and well-structured resources set the standard for how to build dialect speech corpora.

[3] The Nordic Dialect Corpus: http://tekstlab.uio.no/nota/scandiasyn/.

[4] ArchiMob Corpus: https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html.

[5] NECTE Corpus: https://research.ncl.ac.uk/necte/.

In recent years, several spoken and non-standard resources of South Slavic languages have emerged as well. Croatian Adult Spoken Language Corpus (HrAL)[6] (Kuvač and Hržica 2016) samples spontaneous conversations from all Croatian counties, comprising more than 250,000 tokens collected between 2010 and 2016. Bulgarian Dialectology as Living Tradition[7] is a database of oral speech that contains 181 texts representing Bulgarian dialects from 68 different villages across Bulgaria recorded between 1986 and 2013, including short texts from two Torlak villages (Alexander 2015; Zhobov 2011). Bulgarian National Corpus[8] includes over 10 million tokens of transcripts of spoken data from parliament or lectures, but not fully spontaneous speech (Koeva et al 2006). The corpus of spoken Slovene (GOS)[9] (Verdonik and Zwitter Vitez 2011) contains transcripts of speech from various sources, including free conversations from different locations across the country. It is annotated with PoS tags and lemmas, as well as syntactically parsed. Against these resources, Serbian suffers from a lack of a spoken resources that could be used for linguistic analysis and comparison with neighbouring languages.[10]

4 Representing Torlak

Our aim is to create a sample that represents contemporary use of Torlak authentically and accurately, which presupposes a choice and a balance between documenting the non-standard features of this endangered variety and providing the data that will enable an analysis of the variation and the influence of the standard dialect. This involves all steps in the making: in our case, a large amount of recordings of naturally occurring language were collected in the field with a relevant population. A number of representative recordings were then selected for the corpus and transcribed into text for further processing.

4.1 Data collection

Dialect data in the Timok region of Southeast Serbia were collected between 2015 and 2018 within a project called *Čuvari nematerijalne baštine timočkih govora* (Protecting the intangible cultural heritage of the Timok vernacular). The project was oriented towards the preservation of the local language, traditional culture and history. During that time 398 h of audio recordings and 192 h of parallel video recordings were made with more than 200 speakers in 92 villages located in the

[6] HrAL Corpus: https://ca.talkbank.org/access/Croatian.html.

[7] Bulgarian Dialectology Corpus: http://bulgariandialectology.org.

[8] Bulgarian National Corpus: http://search.dcl.bas.bg.

[9] GOS Corpus: http://eng.slovenscina.eu/korpusi/gos.

[10] There exists a corpus of Serbian tweets, which represents colloquial computer-mediated-communication (Miličević and Ljubešić 2016; Miličević et al. 2017).

administrative territories of three municipalities: Knjaževac, Zaječar and Svrljig (for more information on the fieldwork research see Ćirković 2018).[11]

The method of a semi-structured interview was used to elicit long narratives with the goal of capturing spontaneous and naturally occurring speech. Dialect documentation is often based on surveys eliciting some, or only the most exotic structures in a dialect. Nerbonne and Kretzschmar (2013) argue that semi-structured interviews are a better way of collecting speech samples, compared to restricted questionnaires or other unnatural ways of collecting data normally used for atlases. They provide 'data better suited to grammatical analysis and improve researchers' chances of detecting unexpected phenomena' (Nerbonne and Kretzschmar 2013), giving better insight into change and variation.

The interviews were guided but not restricted as to the list of topics referring to history, tradition and culture, folklore, every-day life, biographical stories, agriculture, crafts, and others.[12] The result is a collection of comparable texts which allows inter-speaker comparison. A corpus of comparable texts can be taken as a better ground for comparison because it is not restricted by source language. There are advantages to this approach over the concept of parallel texts which can be influenced by the source language contaminating the translation. Comparable texts control for thematic and, to some extent, linguistic content, while not imposing a formal influence of another language (in favour of the use of comparable texts in contrastive corpus linguistics see Scherrer 2012).

During the fieldwork, it was easy to notice the difference among speakers, and they could roughly be classified in two categories, dialectal and more standard ones. By all means, this distinction is not a binary one but rather gradual. Among the highly non-standard speakers, the large majority are elderly women who have not travelled and which have little or no formal education.[13] Elderly men who had to leave their home for work usually exhibit speech with more SSr elements. The emphasis was on collecting as many samples of the highly non-standard variety, which is considered a rarity, so more effort was invested into reaching older speakers and those who lived in inaccessible locations. Preserving the dialect and knowing more about tradition and history from personal experience made them valuable informants. In order to gain insight into the changes in the language and traditional practices, research included a small number of younger speakers as well.

In most cases, interviews were conducted individually with each speaker to ensure that there were little or no overlaps which could make listening and transcription more difficult. Interviews were usually placed indoors to avoid noise

[11] The project *Čuvari nematerijalne baštine timočkih govora*, funded by the Serbian Ministry of Culture and Information, was active during 2015 and 2016 and consisted in collecting data in the field and publishing online via Timočki govori webpage: http://balksrv2012.sanu.ac.rs/webdict/timok/index. Video recordings collected over the course of research have been published on the Čuvari nematerijalne baštine timočkih govora project webpage and the research team's YouTube channel, *Terenska istraživanja*: https://www.youtube.com/channel/UC4EpCSAnEb2RIsIRY7pfNdQ/feed.

[12] A similar questionnaire has been used in other parts of the Balkans by Sikimić and her team (Sikimić 2012, 2013).

[13] This is a well-known tendency in Slavic dialectology (Belić 1905: XXXIII, Ivić 1985, pp. 92–93), contrary to western dialectology, where male speakers usually preserve the dialect.

Fig. 1 Map of Torlak and Timok area

that could threaten the quality of the recording. Interviewing men and women (usually spouses) together was avoided, because men often took the lead in conversation, being the head of the household, while women would tend to hold back. While conducting the interview, researchers made use of many techniques to make the speakers' production more spontaneous and natural. The common problem of the observer's paradox could be overcome by giving the interviewees more time to relax or initiating more emotionally engaging subject matter, such as important life events, childhood, and the like. In order to avoid eliciting standard forms and to encourage the use of dialect forms, researchers themselves attempted (with varying success) to make use of the dialect.

4.2 Selecting the corpus sample

From the entire collection of the fieldwork recordings a smaller representative sample was selected for the corpus that could be processed with the available resources. The recordings were chosen based on the location and the linguistic production of the speakers, i.e. the level of non-standardness. One criterion was equal geographic coverage of the area. Moreover, ideally, the corpus should comprise a speech sample from each location. This was not always possible, due to the lack or poor quality of recording. The corpus sample includes 80 h of recordings from 64 out of 92 villages in the Timok region. The locations are shown in the map in Fig. 2.

Another criterion was linguistic representativeness. The selection was primarily based on the dialectal markedness of the samples—speakers who use more dialectal features were favoured, including accent, morphological and syntactic features

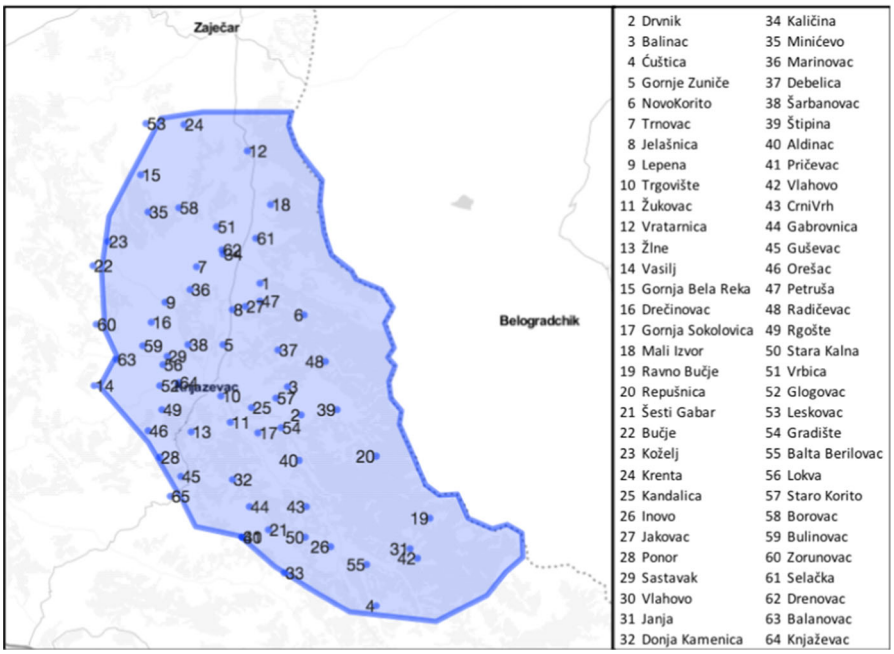

Fig. 2 Map of locations in Timok included in the corpus

regarding verbal and nominal inflection, the use of post-positive demonstrative clitics and others (for relevant features see Vuković and Samardžić 2018). This decision was made by a linguistic expert. It was not always an easy choice, because even the oldest speakers would sometimes use standard forms. As already mentioned, in order to provide data for the analysis of language change, a small number of speakers who speak a variety closer to SSr, including seven high-school students, were interviewed specifically for this purpose, using the same questions (21.07% of the corpus). This part of the corpus does not satisfy the criterion of geographic coverage, since most of these speakers live in the city. These two parts of the corpus are not balanced, as the sample representing the non-standard variety is greater, 68.37% of the entire corpus. The corpus includes the contribution of 12 researchers (10.95% of the text).

Since interviews were in most cases conducted individually, they usually involve only a single speaker. There are rare interviews with more speakers who participate equally in the conversation, while in other interviews some speakers give only brief comments. The total number of participants in the corpus is 166, out of which 80 speakers are present with more than 1000 words each, constituting 80% of the entire corpus.

Table 1 Errors of the initial automatic PoS annotation using the ReLDI tagger for Serbian classified in PoS categories

Noun	Verb	Adj.	Pron.	Adv.	Adp.	Conj.	Num.	Particle	Interj.	Punct.
12.91%	14.53%	1.70%	27.02%	7.31%	2.29%	3.23%	0.68%	27.02%	0.85%	2.38%

Table 2 Lemmatization of non-standard verbs

	Word	Lemma
Torlak	begal run away.M.SG.PPART	begati
SSr	bežao 'run/ran away'	bežati

Table 3 Correction of an unreconstructed lemma

	Word	Automatic lemma	Automatic PoS tag	Corrected lemma	Corrected PoS tag
Tor	juril chase.3SG.PPART	juril	Qo	juriti	Vmp-sm
SSr	jurio 'chased'				

Table 4 Correction of a reconstructed lemma

	Word	Automatic lemma	Automatic PoS tag	Corrected lemma	Corrected PoS tag
Tor	pituju ask.3PL.PRES	pituti	Vmr3p	pitati	Vmr3p
SSr	pitaju '(they) ask'				

4.3 Transcription

Transcripts of fieldwork interviews were made using Partitur-Editor, from the EXMARaLDA software package (Schmidt 2009), which has become a widely used program for transcription. The methodology was based on previous work on oral-dialectal material (Beal et al. 2007) and on the previous experience of the authors

Table 5 Example of the verticalized annotated input file

Word	Lemma	PoS
ja	ja	Pp1-sn
onda	onda	Rgp
uvatim	uhvatiti	Vmr1s
krave	krava	Ncfpa
pa	pa	Cc
vodim	voditi	Vmr1s
		Z

Table 6 Example of the lexicon

Word	Lemma	PoS	Raw freq.	Norm. freq.
kuća	kuća	Ncfsn	6	0.259808
kućata	kuća	Ncfsn-t	3	0.060238
kućava	kuća	Ncfsn-v	2	0.040159
kuće	kuća	Ncfpa	2	0.086603
kuće	kuća	Ncfpn	1	0.043301
kuće	kuća	Ncfsg	1	0.043301
kućete	kuća	Ncfpn-t	1	0.020079

with corpora made up of non-standard spoken language from other varieties of Serbian (Vuković and Miličević 2017; Vuković 2015).

In the transcripts, there are two levels of segmentation, both conveniently enabled by EXMARaLDA. The transcript text is segmented into speakers, by entering the text for each speaker in a separate tier. Within each tier, text was further segmented into speech units. Spoken language is known to have less clear sentence boundaries, making it difficult to divide into sentences such as we find in written texts. Trying to compensate for this, the Timok corpus is broken into segments based on prosody, syntactic patterns and the overall sense. This method had previously successfully been used in the ArchiMob corpus of Swiss dialects (Samardžić et al. 2016). These segments form intonational, structural and meaningful units that provide a frame for the analysis of linguistic patterns. They normally last between 1 and 6 s.

In the transcription text, there is no punctuation nor orthographic capitalization marking sentence structure. The only explicit structural division is created with the so called "event" units in EXMARaLDA, which are partitioned based on time intervals. Capitalization was used to mark accent position (see the passage in Fig. 3). Since corpus search is case-sensitive, the searchable text layer does not show accent position, it is added as a word-level annotation.

baštA ni je bIo • pRvi u • salO • gAzda je bIo || imAo imOvinu •• i svE imAo || nIje kupuvAo nIšta || jedIno je nosIo kom/ krOmpir se mnOgo rađAo || tAmo po ovAj sEla pIrotska || tEmska sOpot #udinje# (Ravno Bučje, 2016, female, 75 years)

'Our father, was the first in the village || he was a landlord, he had property, he had everything || he did not buy anything || he was only carrying pot/ potatoes were growing in abundance || over there in these villages around Pirot || Temska, Sopot, Udinje'

Fig. 3 Transcription in EXMARaLDA

Table 7 The size of the training data

	Verticalized training file	Lexicon
Standard Serbian	88,546	6,908,043
Torlak	26,881	6208
Total	115,427	6,914,251

The transcripts are written in Serbian Latin script, using a semi-phonetic (semi-orthographic) method, which uses a standard alphabet and follows standard orthography while trying to illustrate phonological characteristics of the spoken language and the dialect, a common strategy used in other similar corpora, including FRED corpus of English dialects (Anderwald and Wagner 2007), SCOTS corpus of Scottish varieties (Anderson et al. 2007), as well as transcripts in the corpus of the Bunjevac dialect in Serbia (Vuković and Miličević 2017; Vuković 2015). It creates an easily readable text while preserving information about variability in phonology, which is crucial for non-standard spoken varieties. For example, elided sounds are not reconstructed: the transcript would note *ne mož* instead of the full standard form *ne može* ('cannot').

The transcripts focus on the verbal content of the interviews with the aim to presenting the morpho-phonetic aspects accurately. Normalization was very seldom done directly in the transcripts in case of pronunciation peculiarities or minor errors that are not likely to be repeated in the corpus (e.g. in case of coughing within a word, *br[cough]at*' would be transcribed as *brat ((cough))*, 'brother'). It was important to avoid unnecessary variation and maintain word forms for the sake of diminishing ambiguities and the chance of errors in the further processing. Phonetic changes occurring in word contact were not noted, the words were written to reflect their lexical representation.

Additional remarks in the text refer to pauses, as well as non-verbal voiced elements that can have communicative function, such as laughter, sighs or coughing. Minimal notes of external noise or events (wind, objects in the room, recording devices, etc.) were made when they had an influence on the linguistic production or the sound quality. Unintelligible parts, interrupted words or utterances are marked as well. EXMARaLDA itself provides information about overlapping sections using time-points on the recording. In case of partially unclear words or passages, transcribers tried to reconstruct the content and mark them. An example of a transcript segment within the EXMARaLDA interface is provided in Fig. 3.

Transcription was performed by a team of transcribers and supervised by a team of experts. The experts were members of the research team who collected the data and were closely familiar with the dialect. Each transcript was produced in two stages, the first version was made by a transcriber and then checked and corrected by a supervisor.

For annotation and tool-training purposes, transcript texts were exported in simple text format and labelled in the vertical form (one-token-per-line). In the tokenization, each lexical word was separated as a token, as well as punctuation and other non-lexical content. Pauses and notation for non-verbal sounds (e.g. *((laugh))*, *((cough))*, *((noise))*, etc.) was kept as a separate token, as well as marking of unintelligible sections (*((XXX))*). The mark for interruptions '/' was merged with the word if the word was interrupted in half and separated if the previous word had been spoken entirely. The corpus is stored in TEI (Text Encoding Initiative) conformant XML files, which will be described in more details in the Sect. 7. Transcripts of recordings from Timok comprise 500,697 tokens.

5 Annotation

In order to provide more advanced search options, the corpus includes additional levels of annotation: part-of-speech annotation and lemmatization. Part-of-speech tags with morphosyntactic descriptions (MSD) and lemmatization are an essential part of any corpus, giving information about the grammatical category and the unmarked base form for each token. They are both very useful for inflectional languages because they facilitate search by allowing corpus queries that abstracts away from concrete word-forms, and even more important in case of Torlak, where there is variation in exponents of a single grammatical category. The layer containing information about the accent position (Sect. 4.3) is important, because accent is an important distinguishing dialect feature, that can also be used to predict the distribution of other features (Vuković 2021).

There were no automatic processing tools for annotation and lemmatization previously developed for Torlak. In order to enable the additional search facilities, we are creating tools specifically for this variety. In the case of close varieties, the creation of tools for the under-resourced one often relies on the resourced one for the reasons of efficiency and expedience (see Ljubešić and Klubička 2014; Vuković 2015; Vuković and Miličević 2017). This semi-automatic approach involves annotating a part of the data automatically with the tool for SSr, manually correcting erroneous tags and then using the corrected version to train novel tools. A similar method was applied with other corpora of close varieties (Ljubešić and Klubička 2014; Vuković 2015; Vuković and Miličević 2017). The dialect data was annotated with a tag-set for Serbo-Croatian macrolanguage with adaptations for Torlak and spoken language.

5.1 The Torlak tag-set

The Torlak MSD specifications were defined based on the existing ones for Serbo-Croatian macrolanguage published in the MultextEAST Version 6 (Erjavec 2019). MultextEAST provides language resources and standards for the morphologically rich languages of Central and Eastern Europe (Erjavec 2010, 2012). In the tags, each character position signifies a grammatical category. For example, MSD *Ncms* is

equivalent to the feature-structure consisting of the attribute–value pairs Category=Noun, Type=common, Gender=masculine, Number=singular (Erjavec 2010).

We adapted the set of features included in the tag-set to match the grammar of Torlak. SSr and Torlak differ in morphology and syntax, but there are almost no differences in the inventory of morphological categories. The only grammatical morpheme not contained in the MSD for SSr is the post-posed demonstrative clitic. To include it in the tag-set, an additional character was added at the end of the tag for nouns, adjectives, pronouns and numerals which signifies whether the word hosts a post-positive demonstrative. It can have values*v*, *t*, or *n*, depending on the form of the demonstrative stem,[14] as in the Example (3). The absence of this character position means that the word does not contain the morpheme.

3)		*Unuk-at*	*sadi*	*višnje-te.*
		grandson.M.SG.NOM-DEM	plant.3SG.PRES	cherry.F.PL.ACC-DEM
	PoS tag:	Ncmsnyt	Vmr3s	Ncfpant
		'The grandson is planting the cherries.'		

5.2 Training material for part-of-speech and lemmatization

We have created a training set, for PoS annotation and lemmatization, containing 27,000 tokens. Samples of the data for this purpose were selected from different transcripts to contain many non-standard features from different parts of the area in order to cover a wide range of variation. The data were pre-annotated with ReLDI tagger built for written SSr (Ljubešić et al. 2016a). The tagger assigns an MSD label and a lemma to each word. It applies the MultextEAST MSD labels for Serbo-Croatian macrolanguage. The incorrect lemmas and tags were corrected manually and adapted to the Torlak tag-set where necessary.

5.2.1 Part-of-speech tagging

ReLDI tagger for SSr performed with 63.43% accuracy. Table 1 shows the error rate of the SSr tagger compared to the manually corrected version.

The automatic annotation often made mistakes with dialect forms. Dialect pronoun forms were in many cases labelled with a wrong PoS category. For example, pronouns *tija* and *ovija* (SSr equivalents *taj* and *ovaj*, 'that' and 'this' respectively) were sometimes recognized as an adjective, a noun, or as verb (the latter probably because the form resembles the third person singular of the present tense). Past participle ending in *-l* was often recognized as an adjective, e.g. third person masculine form *bil* of the verb *biti*, 'to be'. Words with non-standard phonemes or phoneme elisions were in many cases assigned an erroneous label with no clear explanation. As expected, the tagger could not recognize words that were chunked due to interruptions.

[14] In Torlak, post-positive demonstrative clitics express three-way deictic reference: v-form (from the demonstrative *ovaj*) signifies a referent close to the speaker, t-form (from the demonstrative *taj*) signifies a referent close to the hearer, while the n-form (from the demonstrative *onaj*) refers to objects far from both the speaker and the hearer.

5.2.2 Lemmatization

The differences in morphosyntax in Torlak have implications for lemmatization. Torlak lacks infinitive forms, like Macedonian and Bulgarian, where lemmas for verbs are given in the first person singular present. Yet as our tools are suited to SSr, they generate infinitive forms as lemmas. Furthermore, as speakers sometimes switch to SSr, infinitive forms in the corpus appear occasionally. For this reason, we decided to preserve infinitives in the lemmatization. In cases of non-standard verbal morphology that affects the stem of the verb to which inflectional suffixes are added, the infinitive was reconstructed, as in the Table 2, even though it is not likely that it would ever be attested in speech.

In cases of morpho-phonetic variants of the same word reflecting diglossia, for instance *medža* and *međa* ('border'), the lemma adheres to the SSr form, as in *međa* in this case. We used the same strategy for other variants, common with pronouns (but in other categories as well). For example, the masculine distal demonstrative pronoun can have two forms in the nominative, the non-standard *onoj*, as well as the SSr *ono* ('that over there'). In a similar fashion, the first person possessive pronoun for feminine objects can have four nominative forms, *ma*, *maa*, *moa* and *moja* ('my'), the fourth being standard. In such cases, the lemma follows from the standard form. For words containing post-positive demonstrative clitics, we ascribe the unmarked nominative form in the lemma. Words that are used with phoneme elisions parallel with their full forms were reconstructed into the full form, as in the case of the adverb *pose* and *posle* ('after').

The accuracy of the automatic lemmatization compared to the manual corrections, was 77.53%. The automatic lemmatizer can reconstruct a form for novel words that are not contained in the lexicon based on the other similar forms and the PoS tag. As expected, errors with non-standard words were common where the PoS tag was wrongly assigned, as in the Table 3. Nevertheless, the automatic lemma reconstruction was sometimes erroneous, even when the PoS tag was correct. This can be observed in the Table 4, where the lemma for the verb contains the infinitive suffix *-ti*, while the stem is incorrect.

5.3 Tagger training

For the automatic PoS annotation and lemmatization, we trained a custom model of the ReLDI tagger (Ljubešić et al. 2016a, b), which is based on the CRF implementation (Okazaki 2007). The tagger requires two input files for training. One of them is a verticalized file containing one token per line with PoS tags and lemmas, with an empty line after each sentence (see Table 5). The other file is a lexicon with one word per line sorted alphabetically, with PoS tags, lemmas and the information about the frequency (raw frequency in the training sample and normalized frequency per 10,000 words), Table 6.

Our training data combines Torlak and Serbian data, which enabled us to have a larger training input by the virtue of the similarity between the two varieties. For the verticalized tagger file, we combined the manually validated Torlak data described above with the manually validated training data for SSr (Batanović et al 2018;

Samardžić et al 2017). We developed the Torlak lexicon based on the dialect training data and added manually annotated instances of words containing the post-positive demonstrative clitics (see Sect. 6.2). The dialect lexicon was merged it with the Serbian lexicon (Ljubešić et al. 2016a, b). The tagger files and instructions are available in a GitHub repository as Torlak-ReLDI-Tagger-2019.[15] The size of the training data and the respective SSr and Torlak sub-sets are given in Table 7 (number of lines in the file).

6 Evaluation

In order to evaluate the quality of the automatic annotation we checked the accuracy on a manually annotated test sample of 2000 words. Apart from estimating the accuracy, the further manual verification gives us an insight into where the errors occur.

Another line of evaluation concerns a corpus query of words containing post-positive demonstratives, given that they are one of the most significant features of the dialect, which also involve the adaptation of the tag-set. This kind of check is important because users are likely to search for dialect forms in the corpus.

6.1 Evaluation of the test sample

We selected 2000 tokens from 2 transcripts for the evaluation of the tagger. Upon manual validation, the accuracy of the tagger is estimated at 84.61% for PoS and 92.62% for lemmatization.

Many errors in the automatic annotation were the same as those described and corrected in the creation of the training set but occurred less frequently. Furthermore, errors often occurred in the case of homonyms, which can be more frequently met in the spoken language. For example, phonological elision of the adverb *sad* has resulted in the form *sa* which is homonymous with the preposition *sa*, and often confused by the tagger. Such errors were not corrected in the release version of the corpus.

Syncretic case forms were often recognised as either in the nominative or accusative form, regardless of their syntactic position. These cases were not taken as erroneous, given that there are no morphological differences between the forms, due do the case loss in Balkan Slavic, and there is not enough surface-level information for the algorithm to distinguish them (consider the difference between *Priča na ženu*. '(Hs) is speaking to a woman.' and *Priča na planinu*. '(He) is speaking on the hill.', both feminine with the same morphological marking, used in a similar context, signifying different syntactic relations, object and location).

[15] Available since 2019 at: https://github.com/bravethea/Torlak-ReLDI-Tagger-2019. Published on CLARIN.SI in 2021 at: http://hdl.handle.net/11356/1378 (Torlak ReLDI Tagger 2021).

6.2 Corpus query: post-positive demonstrative clitics

We searched for the words containing a post-positive demonstrative clitic, like those shown in Example (3). This form can be attached to nouns, pronouns, adjectives and numerals, and as such it is contained in the tag-set.

3)	*Unuk-at*	*sadi*	*višnje-te.*
	grandson.M.SG.NOM-DEM	plant.3SG.PRES	cherry.F.PL.ACC-DEM
PoS tag:	Ncmsnyt	Vmr3s	Ncfpant

'The grandson is planting the cherries.'

On one hand we extracted the words with post-posed demonstratives based on their form: nominal stem with *-at*, *-av*, *-an* clitics and their respective phonological variants and inflected forms distinguishing gender, number and case. On the other hand, we searched for the words using MSD tags: *t*/*v*/*n* in the 6th position for nouns, 7th position in the tag for pronouns and numerals, 8th position in the tags for adjectives.

The first search consisted mostly of manual extraction of the words among other similar forms and returned a total of 1313 words. Out of them, approximately 10% had a correct MSD label and lemma. Out of the correctly annotated words, only 41 were found in the verticalized training data, which means that the tagger could still interpret some novel words. Sometimes, the words were marked correctly for PoS and other grammatical categories, except the post-positive demonstrative. For example, the noun *devojčeto* 'girl', was annotated as *Ncnsn*, but was missing the part related to the post-positive demonstrative, *Ncnsn-t*. Since post-positive demonstratives are a very important characteristic of the dialect, we manually annotated the tags and lemmas for all words carrying the clitic throughout the corpus and included them in the lexicon.

7 Metadata

There are three metadata files, one for the transcripts, one for the speakers and another one for the researchers. The first one keeps the following information:

- ID: a unique ID for each recording-transcript pair
- Location: the location that the speaker represents
- Longitude: longitude from Google Maps
- Latitude: latitude from Google Maps
- Transcriber: full names of the transcribers
- REC_duration: the duration of the source recording in the hh:mm:ss format
- Tokens: the number of tokens in the transcript
- Utterances: the number of segments in the transcript
- Researchers: full names of the researchers conducting the interview

The speaker metadata file contains the following information:

– ID: a unique ID for each speaker
– Location: place of residence
– Origin: birth-place, if different from residence
– Gender
– Year of birth
– Age
– Education
– Occupation
– Dialect speaker, yes/no: whether they are representative for the dialect (judged by an expert)

The researchers metadata file contains the following information:

– ID: a unique ID for each researcher
– Name: Full name
– Gender.

Many researchers involved in the fieldwork come from an anthropological and ethnographic background where regular fieldwork practice does not give importance to speaker metadata, which is extremely important for linguistic research. Information about individual speakers is thus unknown. In some cases, it can be guessed or approximated based on their appearance and the content of the interview. These figures were marked specifically and can be excluded. In other cases, this information is utterly missing. No metadata on researchers were recorded, as it is not the focus of the research and does not impact the dialectal content of the interview.

8 Corpus files and corpus access

The data are stored in several types of files as listed below. The TEI files follow the Standard for Transcriptions of Spoken Language[16] (ISO 24624 2016):

– TEI XML corpus files containing the transcripts, together with annotation, and the information about the projects
– CSV Metadata files
– TEI XML file with the PoS tagset definitions
– EXMARaLDA transcription files
– TXT files extracted for each transcript and for each speaker

The corpus XML files start with a header containing information about the transcript, a timeline element with the time intervals and the corpus text. The structure of the XML files is sketched out in the Example 4. The element '*u*' contains an utterance or a chunk of transcribed speech with the information about time-alignment and the speaker ID. Inside the element '*u*', '*w*' is for words, keeping the additional information (lemma, MSD tag ('*ana*')), '*pause*' is for silent pauses,

[16] Available at: https://www.iso.org/standard/37338.html.

'*vocal*' is for laughter, coughing, filled pauses and other human non-verbal sounds, '*incident*' is for non-human sounds. Sections that have not been transcribed due to unintelligibility are represented under the '*unclear*' element, while those that have been omitted for being irrelevant or sensitive are marked as '*gap*'.

4)

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="TOR_C_0001" xml:lang="sr">
  <teiHeader>
[...]
<text>
    <body>
        <timeline unit="s" origin="#TOR_C_0001.t0" corresp="TOR_C_0001.wav">
        <when xml:id="TOR_C_0001.t0"/>
        <when xml:id="TOR_C_0001.t21" interval="645.8259606710163" since="#TOR_C_0001.t0"/>
        <when xml:id="TOR_C_0001.t22" interval="646.1659602993401" since="#TOR_C_0001.t0"/>
         […]
        </timeline>
        <u xml:id="TOR_C_0001-u12" start="TOR_C_0001.T21" end="TOR_C_0001.T22" who="#TIM_SPK_0001">
            <w xml:id="TOR_C_0001-u433-w1" lemma="trčati" ana="mte:Vmr1p">trčimó</w>
            <w xml:id="TOR_C_0001-u433-w2" lemma="devočići" ana="mte:Qr">devóčići</w>
            <pause xml:id="TOR_C_0001-u434-w7"/>
            <vocal xml:id="TOR_C_0001-u522-w1" type="laugh"/>
            <incident xml:id="TOR_C_0015-u51-w1" type="noise"/>
            <unclear/>
            <gap reason="editorial"/>
        </u>
    </body>
    </text>
    </TEI>
```

The corpus, together with the tools and training sets will be published within the CLARIN.SI platform for language resources[17] under the Creative Commons Attribution-NonCommercial 4.0 licence. The repository comprises many resources for various languages, focusing on South Slavic. The corpus will also be made available through the NoSketch and the KonText platforms. Audio files in WAV format, in which the untranscribed sections have been masked, are being prepared for publication separately under the same licence in 2021. The corpus text and the recordings cannot at the moment be searched in parallel using a corpus search platform, but they can be viewed and searched using the EXMARaLDA software.

Through the NoSketch interface, it is possible to search the corpus text and annotations (individually or combined) using regular expressions and Corpus Query Language (Jakubíček et al. 2010). The linguistic information extracted from the text can be combined and correlated with the extralinguistic information contained as metadata. Users can also download the corpus files and search or process it locally.

9 Discussion

The corpus of Torlak aims to provide previously missing data for the study of Torlak that could prove useful for low-resourced natural language processing. At the same time, the corpus documents an endangered variety, capturing what was probably the last opportunity to note some exotic features used only by the elderly.

Owing to the fact that linguistic data of speakers of different age was sampled, the corpus allows for the analysis of language change over the last few decades. As Mandić (2016) points out, Serbian dialectology has been oriented towards rural

[17] Available at: http://hdl.handle.net/11356/1281.

vernaculars and descriptions of exclusively non-standard features. Linguists often neglect the degrees of change between non-standard/rural and standard/urban varieties—a drawback that can now be remedied with the corpus at hand.

Using the MultextEast PoS annotation makes the data comparable with the other neighbouring languages of Southeast Europe. The labels can be easily transformed into today's widely used standard, Universal Dependencies (Zeman et al. 2019), since they involve overlapping morpho-syntactic categories. This can be seen in the manually verified sample for Serbian, SETimes.SR 1.0, where both sets of labels are included (Batanović et al. 2018; Samardžić et al. 2017).

The semi-orthographic transcripts preserve information about the non-standard phonology and the peculiarities of the spoken language while making the texts easily readable for humans and machines. The transcripts are not always a verbatim representation of the recordings, as described in the Sect. 4.3, and involve normalizations of word-final sound changes as well as rare incidental word chunking. Given such normalizations, the chosen approach might not be ideal for a potential contribution to automatic speech recognition; nonetheless, it is far from useless, considering that the corpus contains a large collection of time-aligned text.

The first release version of the corpus represents the dialect sample from Timok. In the next release, data from other Torlak regions will be added. Interviews using the same methodology have also been collected in the Lužnica region in Serbia, as well as the areas around Tran and Belogradchik in Bulgaria. In order to enable diachronic analysis of language change, digitized transcripts from Sobolev's Atlas (Sobolev 1998) will be included as well. The team also strives to improve the automatic tools used for the corpus with the future releases.

There has been an attempt to add morphological and syntactic normalization into SSr to the corpus, which would have allowed users to access non-standard forms through their standard equivalents. However, the results of the automatic normalization using were not satisfying, which is why this remains a task for a future release of the corpus.

This may be important when it comes to one commonly used approach for developing annotation tools for un-resourced languages, which relies on normalization or a parallel text in a resourced language (cf. Zennaki et al. 2015; Vuković and Miličević 2017). The pipeline involves normalizing non-standard data into the respective standard variety, annotating the normalized version of the text with an existing tagger, and then projecting annotations onto the original text. This approach is convenient because it helps process under-resourced text using existing tools and adds an additional level of normalization. Nevertheless, we chose to annotate the data directly for two reasons. Firstly, the MSD in SSr and Torlak do not always overlap. For instance, the suppletive case forms in SSr typically marked as locative should only be marked as accusative/oblique in Torlak (according to linguists' descriptions, e.g. Belić 1905; Tomić 2006). Secondly, and more importantly, the final version of the corpus should include data from many sources, which creates a challenge for normalization into a single variety, especially in regard to the data from Bulgaria. We therefore chose to build a tagger specifically for Torlak that could be used for the whole corpus at a later stage.

The data form the corpus has already been used in several studies focusing on the diffusion of linguistic features in the area, as well as linguistic variation.[18] Apart from linguistics, the data can be used for various other topics—anthropology, ethnography, history, sociology, among others. An additional advantage comes from the fact that similar questionnaires have been used by researchers doing fieldwork in Timok, Serbia and the Balkans for decades. This makes this collection potentially comparable with transcripts of surrounding varieties, should they become electronically available in the future (see Sikimić 2012; Bošnjaković and Sikimić 2013).

10 Conclusion

The corpus of the Torlak dialect spoken in the Timok area in Southeast Serbia documents an endangered variety and provides valuable research material of spoken and non-standard Serbian. The sample comprises 80 h of field interviews collected between 2015 and 2017. Data include semi-orthographic transcripts from 64 villages, with accent information, annotated with PoS labels and lemmatization. The region is covered geographically, and the sample represents people of different age groupings. The metadata contains demographic and geographic information which can be used to analyse correlations between the linguistic content and extralinguistic factors. The scope of the topics covered in the interviews makes the corpus useful for different disciplines, including anthropology, history, sociology, and others. Upon examining the accuracy of the automatic tagger, some annotations that were important for the representation of the dialect were corrected throughout the corpus. The corpus is made freely available along with the tools for automatic annotation and manually verified training sets.

Acknowledgements Data collection and corpus creation were supported by multiple funding parties. Fieldwork was conducted within the project *Čuvari nematerijalne baštine timočkih govora* (Protecting the intangible cultural heritage of the Timok vernacular) funded by the Ministry of Culture and Information of the Republic of Serbia in 2015 and 2016. Transcription and further processing of data were funded by the Language and Space laboratory, Slavisches Seminar, Doctoral Program in Linguistics, and the SyNoDe project of the University of Zurich. The process of corpus creation and data analysis continues within the TraCeBa project, funded by the Swiss National Science Foundation, the Russian Academy of Sciences and Ministry of Education, Science and Technological Development of the Republic of Serbia through the EraNet Rus Plus funding scheme. My PhD research has also been funded by the Swiss Government Excellence Scholarship. I would like to express my gratitude to the team of fieldwork researchers on their efforts to find as many dialect speakers as possible, reaching even the most inaccessible locations; to transcribers on hours spent converting recordings into text; to Sanja Bradjan and Danijela Stojković for manual annotations, Dr. Tanja Samardžić and Dr. Nikola Ljubešić for their NLP expertise and help with data processing, and last, but certainly not least, to my supervisors Prof. Dr. Barbara Sonnenhauser, Dr. Hanne Eckhoff for supporting my PhD research.

Author contributions Writing the paper; collecting and processing the data for the corpus.

[18] References will be added in the final version, because they all contain the name of the author of the present manuscript, which is why they have not been mentioned in the versions under review.

Funding Open Access funding provided by University of Zurich. Funding was provided by ERA-Net Rus Plus (Grant No. 307), Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (CH) (Grant No. 177557), Swiss Government Excellence Scholarship, Ministry of Education, Science and TechnologicalDevelopment, Republic of Serbia.

Data availability The materials are made available as described in the text.

Code availability The tagger code is made available on the link provided in the text.

Compliance with ethical standards

Conflict of interest The author declare that they have no conflict of interest.

Informed consent All the participants in the corpus have given their consent to be a part of the research and for the data to be published.

References

Alexander, R. (1975). *Torlak accentuation*. Munich: Verlag Otto Sagner.

Alexander, R. (2015). Bulgarian dialectology as living tradition: A digital resource of dialectal speech. *Balkanistica 28* (pp. 1–13). Pennsylvania State University Press.

Allen, W. H., Beal, J. C., Corrigan, K. P., Maguire, W., & Moisl, H. L. (2007). A linguistictime-capsule: The Newcastle Electronic Corpus of Tyneside English. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitising language corpora, Vol. 2: Diachronic databases* (pp. 16–48). Houndmills: Palgrave Macmillan.

Anderson, J., Beavan, D., & Kay, C. (2007). SCOTS: Scottish corpus of texts and speech. In J. Beal, K. Corrigan, & H. Moisl (Eds.), *Creating and digitizing language corpora. Volume 1: Synchronic databases* (pp. 17–34). Basingstoke: Palgrave Macmillan.

Anderwald, L., & Wagner, S. (2007). FRED—The Freiburg English Dialect Corpus: Applying corpus-linguistic research tools to the analysis of dialect data. In J. Beal, K. Corrigan, & H. Moisl (Eds.), *Creating and digitizing language corpora. Volume 1: Synchronic databases* (pp. 35–53). Basingstoke: Palgrave Macmillan.

Batanović, V., Ljubešić, N, Samardžić, T., & Erjavec, T. (2018). Training corpus SETimes.SR 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1200.

Beal, J., Corrigan, K., & Moisl, H. (Eds.). (2007). *Creating and digitizing language corpora. Volume 1: Synchronic databases*. Basingstoke: Palgrave Macmillan.

Belić, A. (1905). *Dijalekti istošne i južne Srbije*. Beograd: Srpska kraljevska akademija.

Bogdanović, N. (2015). *Govor i jezik Staroplaninaca*. Niš: Galakianis.

Bošnjaković Ž., & Sikimić, B. (2013). Bunjevci. Etnodijalektološka istraživanja 2009. Novi Sad: Matica srpska. Subotica: Nacionalni savet bunjevačke nacionalne manjine.

Ćirković, S. (Ed.). (2018). *Timok. Folkloristička i lingvistička terenska istraživanja 2015–2017, Knjaževac: Narodna biblioteka "Njegoš"* (pp. 219–238). Beograd: Udruženje folklorista Srbije.

Dinić, J. (2008). *Timočki dijalekatski rečni*. Beograd: Institut za srpski jezik SANU.

Erjavec, T. (2010). MULTEXT-East version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the LREC 2010, Malta, 19–21 May, 2010*. Malta: ELRA.

Erjavec, T. (2012). MULTEXT-East: Morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation, 46*(1), 131–142.

Erjavec, T. (2019). Serbo-Croatian specifications. In Erjavec, T. (Ed.), MULTEXT-East morphosyntactic specifications. http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html. Accessed on 15 December 2019.

Graul, C. (2016). leafletR: Interactive web-maps based on the Leaflet JavaScript Library. R package version 0.4-0. http://cran.r-project.org/package=leafletR.

Ivić, P. (1985). *Dijalektologija srpskohrvatskog jezika. Uvod i štokavsko narečje*. Novi Sad: Matica srpska.

ISO 24624. (2016). *Language resource management—Transcription of spoken language*. ISO. https://www.iso.org/standard/37338.html.

Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. In *PACLIC*, pp. 741–747.

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A., & Vangsnes, Ø. A. (2009). The Nordic Dialect Corpus—An advanced research tool. In K. Jokinen, & E. Bick (Eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. NEALT Proceedings Series Volume 4.

Johannessen, J. B., Vangsnes, Ø. A., Priestley, J., & Hagen K. (2014). A multilingual speech corpus of North-Germanic languages. In T. Raso, & H. Mello (Eds.), *Spoken corpora and linguistic studies* (pp. 69–83). John Benjamins Publishing Company.

Koeva, S., Leseva, S., Stoyanova, I., Tarpomanova, E., & Todorova, M. (2006). Bulgarian tagged corpora. In *Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages* (pp. 78–86). Sofia, Bulgaria.

Kuvač, J. K., & Hržica, G. (2016). Croatian adult spoken language corpus (HrAL). In *Fluminensia: Journal for Philological Research, 28*, 2. Rijeka: University of Rijeka, Faculty of Humanities and Social Sciences.

Lindstedt, J. (2000). Linguistic Balkanization: Contact-induced change by mutual reinforcement. In D. Gilbers, J. Nerbonne, & J. Schaeken (Eds.), *Languages in contact*. Amsterdam & Atlanta: Rodopi.

Ljubešić, N., & Klubička, F. (2014). {bs,hr,sr}WaC—Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)* (pp. 29–35). Gothenburg: Association for Computational Linguistics.

Ljubešić, N., Klubička, F., Agić, Ž., & Jazbec, I. P. (2016a). New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: ELRA.

Ljubešić, N., Klubička, F., & Boras, D. (2016b). Inflectional lexicon srLex 1.2. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1073.

Mandić, M. (2016). Vernakularna autentičnost u srpskoj dijalektologiji: kritičko preisipitivanje. In R. Žugić (Ed.). *Dijalekti Srpskoga Jezika: Istraživanja, Nastava, Književnost* (pp. 519–530). Leskovac. Leskovački kulturni centar. Vranje: Univerzitet u Nišu, Pedagoški fakultet.

Miličević, M., & Ljubešić, N. (2016). Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research, 4*(2), 156–188.

Miličević, M. P., Ljubešić, N., & Fišer, D. (2017). Nestandardno zapisivanje srpskog jezika na Tviteru: mnogo buke oko malo odstupanja?. In *Anali Filološkog fakulteta* (Vol. 29, No. 2, pp. 111–136). Belgrade: Filološki fakultet.

Nerbonne, J., & Kretzschmar, W., Jr. (2013). Dialectometry ++. *LLC: Journal of Digital Scholarship in the Humanities, 28*(1), 2–12.

Okazaki, N. (2007). CRFsuite: A fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Penev, G., & Marinković, I. (2012). Prvi rezultati popisa stanovništva 2011. S posebnim osvrtom na promenu broja stanovnika jugoistočne Srbije. Stanovništvo jugoistočne Srbije: uticaj demografskih

promena u jugoistočnoj Srbiji na društveni razvoj i bezbednost, Niš: Centar za naučnoistraživački rad SANU i Univerziteta u Nišu, 21–42.
Petrović, T. (2015). *Srbija i njen jug. "Južnjački dijalekti" između jezika, kulture i politike*. Beograd: Fabrika knjiga.
Salminen, T. (2010). Europe and Caucasus. In C. Moseley (Ed.), *Atlas of the World's languages in danger* (pp. 32–42). Paris: UNESCO Publishing.
Samardžić, T., Scherrer, Y., & Glaser, E. (2016). ArchiMob—A corpus of spoken Swiss German. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, & B. Maegaard et al. (Eds.), *Language resources and evaluation (LREC 2016)* (pp. 4061–4066). Portorož, Slovenia.
Samardžić, T., Starović, M., Agić, Ž., & Ljubešić, N. (2017). Universal dependencies for Serbian in comparison with Croatian and Other Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (pp. 39–44). Valencia: ACL.
Scherrer, Y. (2012). Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH (EACL 2012)* (pp. 63–71). Stroudsburg, PA: Association for Computational Linguistics.
Scherrer, Y., Samardžić, T., & Glaser, E. (2019). Digitising Swiss German: How to process and study a polycentric spoken language. *Language Resources and Evaluation, 53,* 735–769.
Schmidt, T. (2009). Creating and working with spoken language corpora in EXMARaLDA. In V. Lyding (Ed.), *Lesser used languages & computer linguistics II* (pp. 151–164). Eurac Research: Bolzano.
Schmidt, T. (2014). The database for Spoken German—DGD2. In *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).
Sikimić, B. (2012). Timski terenski rad Balkanološkog instituta SANU. Razvoj istraživačkih ciljeva i metoda. In M. Ivanović-Baričić (Ed.), *Terenska istraživanja—poetika susreta* (pp. 167–198). Belgrade: The Institute of Ethnography SASA.
Sikimić, B. (2013). Pergled terenskin istraživanja Srba u Mađarskoj obavljenih u okviru Balkanološ kog instituta SANU, Beograd (2001–2010). In P. Lastić, & D. Radojičić (Eds.), *Etnologija Srba u Mađarskoj: stanje i perspektive* (pp. 93–102). Budapest: Srpski institut. Belgrade: Etnografski institut SANU.
Sobolev, A. N. (1998). *Sprachatlas Ostserbiens und Westbulgariens*. Marburg/Lahn: Biblion-Verlag.
Stanojević, M. (1911). *Severno timočki dijalekat*. Beograd: Srpska kraljevska akademija.
Tomić, O. M. (2006). *Balkan Sprachbund morpho-syntactic features.* Studies in natural language and linguistic theory 67. Dordrecht: Springer.
Verdonik, D., & Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos*. Ljubljana: Trojina.
Vuković, T. (2015). Izrada modela dijalekatskog korpusa bunjevačkog govora. Beograd: Filološki Fakultet (unpublished MA thesis). https://www.academia.edu/20622305/Izrada_modela_dijalekatskog_korpusa_bunjevačkog_govora. Accessed 25 March 2019.
Vuković, T. (2021). Torlak ReLDI Tagger 1.0. CLARIN.SI. http://hdl.handle.net/11356/1378.
Vuković, T., Escher, A., & Sonnenhauser, B. (submitted). Degrees of non-standardness. Feature-based analysis of variation in a Torlak dialect corpus.
Vuković, T., & Miličević, M. (2017). Creation and some ideas for classroom use of an electronic corpus of the dialect of Bunjevci. u: J. Filipović and J. Vučo (Eds.), *Minority. Languages in education and language learning: Challenges and new perspectives*. Belgrade: Faculty of Philology.
Vuković, T., & Samardžić, T. (2018). Prostorna raspodela frekvencije postpozitivnog člana u timočkom govoru. In S. Ćirković (Ed.), *Timok: folkloristička i lingvistička terenska istraživanja 2015–2017*. Knjaževac: Narodna biblioteka Njegoš.
Woolhiser, C. (2011). Border effects in European dialect continua: Dialect divergence and convergence. In B. Kortmann & J. van der Auwera (Eds.), *The languages and linguistics of Europe: A comprehensive guide* (pp. 501–523). Berlin: Walter de Gruyter.
Zeman, D., Nivre, J., Abrams, M., et al. (2019). Universal dependencies 2.5, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3105.
Zennaki, O., Semmar, N., & Besacier, L. (2015). Unsupervised and lightly supervised part-of-speech tagging using recurrent neural networks. *29th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, Shangai, China: The 29th Pacific Asia Conference on Language, Information and Computation.

Zhobov, V. (2011). Проект за дигитално представяне на българските диалекти-Компютърни и интерактивни средства за исторически езиковедски изследвания/A project for the digital representation of Bulgarian dialects: Computerized interactive means for historical linguistics. In *Сборник доклади от заключителна конференция/Papers from the Closing Conference. Sofia* (pp. 28–35).